Wavelets, Approximation and Compression: A Review

Martin Vetterli

Laboratory for AudioVisual Communications,
Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland

Dept. of Electrical Engineering and Computer Sciences,
University of California, Berkeley, CA 94720

ABSTRACT

In this paper, we briefly review the connection between subband coding, wavelet approximation and general compression problems. Wavelet or subband coding is successful in compression applications partly because of the good approximation properties of wavelets. First, we revisit some rate-distortion bounds for wavelet approximation of piecewise smooth functions. We contrast these results with rate-distortion bounds achievable using oracle based methods. We indicate that such bounds are achievable in practice using dynamic programming. Finally, we conclude with an outlook on open questions in the area of compression and representations.

Keywords: subband coding, wavelet coding, approximation, rate-distortion

1. INTRODUCTION

Subband coding is a speech compression technique dating from the late 70's,[5] and was generalized to images[19,21] and video[8] in the 80's. The introduction of wavelets[6] was soon followed by the recognition that they have a close connection to subband coding.[9] By now, subband coding of images using a wavelet-like decomposition are very popular.[17] As a measure of success of these methods, we can mention that the JPEG-2000 image compression standard (currently in the making) will very likely be based on wavelet decompositions.

So, are wavelets a magic tool, a silver bullet for a whole range of problems, from approximation to compression?

The answer is certainly not so simple. Clearly, an alternative signal representation to local Fourier analysis was long overdue, and wavelets are such an alternative. In particular, they are able to perform both a local and a global analysis of a signal. Thus, wavelets are well suited for the analysis of piecewise smooth signals, for which Fourier bases are lacking.

Let us briefly recall the approximation properties of wavelet bases, concentrating on the orthonormal case for simplicity.[18,10] Given a wavelet with N zero moments (or N zeros at $\omega = 0$) then

1. The wavelet is of length at least $2N - 1$.
2. Polynomials up to degree $N - 1$ have vanishing inner products with the wavelet at all scales.
3. Polynomials up to degree $N - 1$ are reproduced by linear combination of the scaling function.
4. At singular points (e.g. from one polynomial piece to the next), at most $2N - 1$ wavelet coefficients differ from zero at each scale.

As will be reviewed below, this leads to efficient non-linear approximation schemes for piecewise polynomial functions in wavelet bases.

However, such non-linear approximation power is only one side of the compression tapestry, the other one being the bitrate consumed by this non-linear approximation scheme. Therefore, in the section below, we review the compression problem in detail, showing how it lives at the intersection of several fields.

2. THE COMPRESSION PROBLEM

Let us review what is the kernel problem in compression. Compression is the trade-off between description complexity and approximation quality. Given an object of interest, or a class of objects, one studies this trade-off by choosing a representation (e.g. an orthonormal basis) and then deciding how to describe the object parsimonously in the representation. Such a parsimonous representation typically involves approximation.

For a function described with respect to an orthonormal basis, only a subset of basis vectors might be used (subspace approximation) and the coefficients used in the expansion are approximated (quantization of the coefficients). Thus, both the subspace approximation and the coefficient quantization contribute to the approximation error. More formally, for a function f in $L_2(R)$ for which $\{\varphi_n\}$ is an orthonormal basis, we have the approximate representation:

$$\hat{f} = \sum_{n \in I} \hat{\alpha} \varphi_n \tag{1}$$

$$\hat{\alpha}_n = Q[< \varphi_n, f >] \tag{2}$$

where I is an index subset and $Q[.]$ is a quantization function, like for example the rounding to the nearest multiple of a quantization step $\triangle$:

$$\hat{\alpha} = Q[\alpha] = \triangle \cdot \left(\left\lfloor \frac{\alpha}{\triangle} \right\rfloor + \frac{1}{2} \right). \tag{3}$$

Usually, the approximation error is measured by the L_2 norm, or squared distortion

$$\epsilon = \|f - \hat{f}\|_2^2 \tag{4}$$

While there is a lot of debate about the relevance of the L_2 norm for practical schemes, e.g. where human viewers are involved, it is still the most commonly used norm, and one of the few for which quantitative results are possible.

The description complexity corresponds to describing the index set I, as well as describing the quantized coefficients $\hat{\alpha}_n$. The description complexity is usually called the rate R, corresponding to the number of binary digits (or bits) used.

Therefore the approximation $\hat{f}$ of f leads to a rate-distortion pair (R, ϵ), indicating one possible trade-off between description complexity and approximation error.

The example just given, despite its simplicity, is quite powerful and actually used in practical compression standards.

It also raises the following questions[7]:

Q1: What are classes of objects of interest and for which the rate-distortion trade-off can be well understood?

Q2: If approximations are done in bases, what are good bases to use?

Q3: How to choose the index set and the quantization?

Q4: Are there objects for which approximation in bases is suboptimal?

Historically, Q1 has been addressed by the information theory community in the context of rate-distortion theory. Shannon posed the problem in his 1948 landmark paper.[15] and proved rate-distortion results in his 1959 paper.[16] The classic book by Berger[1] is still a reference on the topic. Yet, rate-distortion theory has been mostly concerned with exact results within an asymptotic framework (the so-called large blocksize assumption together with random coding arguments). Thus, only particular processes (e.g. jointly Gaussian processes) are amenable to this exact analysis. But the framework has been used extensively, in particular in its operational version (when practical schemes are involved), see for example.[12] It is to be noted that rate-distortion analysis covers all cases (e.g. small rates with large distortions) and that the case of very fine approximation (or very large rates) is usually easier but less useful in practice.

The second question has a simple answer, based on rate-distortion theory, in the stationary jointly Gaussian case. Then, the canonical basis is the Karhunen-Loève basis, and a procedure called reverse waterfilling leads to the optimal behavior. Yet, not all things in life are jointly Gaussian, and that is where wavelets come into play. For processes which are piecewise smooth (e.g. images), the abrupt changes are well captured by wavelets, and the smooth or stationary parts are efficiently represented by coarse approximations using scaling functions. Both practical algorithms (e.g. the EZW algorithm of Shapiro[17] and theoretical analysis thereof[3,11]) have shown the power of approximation within a wavelet basis.

An alternative is to search large libraries of orthonormal bases, based for example on binary subband coding trees. This leads to wavelet packets[4] and algorithms for rate-distortion optimal solutions.[14]

The third question is more complex than it looks at first sight. If there was no cost associated with describing the index set, then clearly I should be the set $\{n\}$ such that

$$| < \varphi_n, f >_{n \in I} | \geq | < \varphi_m, f > |_{m \notin I} \tag{5}$$

But when the rate for I is accounted for, it might be more efficient to use a fixed set I for a class of objects.

For example, in the jointly Gaussian case, the optimal procedure chooses a fixed set of Karhunen-Loève basis vectors (namely those corresponding to the largest eigenvalues) and spends all the rate to describe the coefficients with respect to these vectors. Note that a fixed subset corresponds to a linear approximation procedure (before quantization, which is itself non-linear) while choosing a subset as in (5) is a non-linear approximation method.

It is easy to come up with examples of objects for which non-linear approximation is far superior to linear approximation. Consider a step function on $[0,1]$, where the step location is uniformly distributed on $[0,1]$. Take the Haar wavelet basis as an orthonormal basis for $[0,1]$. It can be verified that the approximation error using M terms is of the order

$$\epsilon_L \sim 1/M \tag{6}$$

for the linear case, while it is of the order

$$\epsilon_{NL} \sim 2^{-M} \tag{7}$$

for non-linear approximation using the M largest terms. However, this is only the first part of the rate distortion story, since we still have to describe the M chosen terms.

This rate distortion analysis takes into account that a certain number of scales J have to be represented, and at each scale, the coefficients require a certain number of bits. This split leads to a number of scales $J \sim \sqrt{R}$. The error is the sum of errors of each scale, each of which is of the order $2^{-R/J}$. Together, we get:

$$D_{NL}(R) \sim \sqrt{c \cdot R} 2^{-\sqrt{c \cdot R}} \tag{8}$$

The quantization question is relatively simple if each coefficient α_n is quantized by itself (so-called scalar quantization). Quantizing several coefficients together (or vector quantization) improves the performance, but increases complexity. Usually, if a "good" basis is used and complexity is an issue, scalar quantization is the preferred method.

The fourth question is a critical one. While approximation in orthonormal bases are very popular, they cannot be the end of the story. Just as not every stochastic process is gaussian, not all objects will be well represented in an orthonormal basis. In other words, fitting a linear subspace to arbitrary objects is not always a good approximation. But even for objects where basis approximation does well, some other approximation method might do much better. In our step function example studied earlier, a simple minded coding of the step location and the step value leads to a rate-distortion behavior

$$D'(R) \sim 2^{-c' \cdot R} \tag{9}$$

The difference between wavelet and direct approximation of piecewise polynomial signals is investigated in[13] to which we refer for further details.

3. DISCUSSION

The above example shows that a generic compression scheme based on wavelets can be quite suboptimal when it comes to a specific class of objects like piecewise polynomial functions. So while wavelet schemes do much better than a global Fourier method (where the non-linear approximation behavior would be quite poor because of the singular points), they are not the ultimate answer either.

Another topic of importance is the generalization to higher dimensions, e.g. for two-dimensional functions like images. There, piecewise smooth means singularities like edges, ridges etc. These are singular in one direction (across the edge) but regular in the other (along the edge). Now, two-dimensional wavelets which are essentially cross-products of one-dimensional ones are suited for point singularities, but not edge singularities. Thus, new constructions are necessary, like for example ridgelets.[2] However, rate-distortion behaviors of such schemes are not known, and thus, many questions remain open for compression of two-dimensional objects.

REFERENCES

1. T. Berger, **Rate Distortion Theory**, Prentice-Hall, Englewood Cliffs, NJ, 1971.
2. E.Candes and D.Donoho, "Ridgelets: a key to higher-dimensional intermittency?," Phil. Trans. R. Soc. Lond. A. to appear.
3. Cohen, A., Daubechies, I., Guleryuz, O.,and Orchard, M., "On the importance of combining wavelet-based non-linear approximation in coding strategies, Manuscript, 1997.
4. R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection", IEEE Tr. on IT, Vol. 38, No. 2, pp. 713-718, March 1992.
5. A. Croisier and D. Esteban and C. Galand, "Perfect Channel Splitting by Use of Interpolation/Decimation/Tree Decomposition Techniques," Int. Conf. on Inform. Sciences and Systems Aug. 1976, pp. 443–446, Patras, Greece.
6. I. Daubechies, "Orthonormal Bases of Compactly Supported Wavelets," Commun. on Pure and Appl. Math., Nov. 1988, Vol. 41, pp. 909-996.
7. D.Donoho, M.Vetterli, R.DeVore and I.Daubechies, "Data Compression and Harmonic Analysis," IEEE Tr. on IT, Special Issue, Information Theory: 1948-1998 Commemorative Issue, Vol. 44, No. 6, Oct.1998, pp.2435-2476.
8. G. Karlsson and M. Vetterli, "Sub-band coding of video signal for packet-switched networks," Proc. of SPIE Conf. on Visual Communications and Image Processing, Vol. 845, Cambridge, MA, Oct. 1987, pp. 446-456.
9. S. Mallat, "A Theory for Multiresolution Signal Decomposition: the Wavelet Representation," IEEE Tr. on PAMI, Vol. 11, No. 7, July 1989, pp. 674-693.
10. S.Mallat, **A Wavelet Tour of Signal Processing**, Academic Press, San Diego, 1998.
11. S.Mallat and F.Falzon, "Analysis of low bit rate image transform coding," *IEEE Tr. SP*, **46**, no.4, pp 1027-42, April 1998.
12. A.Ortega and K.Ramchandran, "Rate-distortion methods for image and video compression," IEEE Signal Processing magazine, Oct. 1998.
13. P.Prandoni and M.Vetterli, "Approximation and compression: rate-distortion bounds for wavelet and direct schemes," Phil. Trans. R. Soc. Lond. A. to appear.
14. K.Ramchandran and M.Vetterli, "Best wavelet packet bases in a rate-distortion sense," IEEE Trans. on Image Processing, Vol. 2, No. 2, April 1993, pp. 160-175.
15. C. E. Shannon, "A mathematical theory of communication," Bell Syst. Tech. Journal, Vol.27, 1948.
16. C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," IRE Nat. Conv. Rec. 1959, pp. 142-163.
17. J. M. Shapiro, "Embedded Image Coding Using Zerotrees of Wavelet Coefficients," IEEE Tr. on Signal Processing, Vol. 41, No. 12, Dec. 1993, pp. 3445–3462.
18. G. Strang and T. Nguyen, **Wavelets and Filter Banks**, Wellesley Cambridge Press, 1996.
19. M.Vetterli, "Multi-dimensional subband coding: some theory and algorithms," Signal Processing, Vol. 6, No.2, pp. 97-112, April 1984.
20. M.Vetterli and J.Kovačević, **Wavelets and Subband Coding**, Prentice-Hall, Englewood Cliffs, NJ, 1995.
21. J. W. Woods and S. D. O'Neil, "Sub-Band Coding of Images," IEEE Tr. on ASSP, Vol. 34, No. 5, pp. 1278-1288, May 1986.